Climate Data Analysis Tools (CDAT):

Enabling data management, data analysis, and visualization for climate scientific discovery

Quarterly Progress Report for the Period

November 2005 through January 2006

1. Introduction

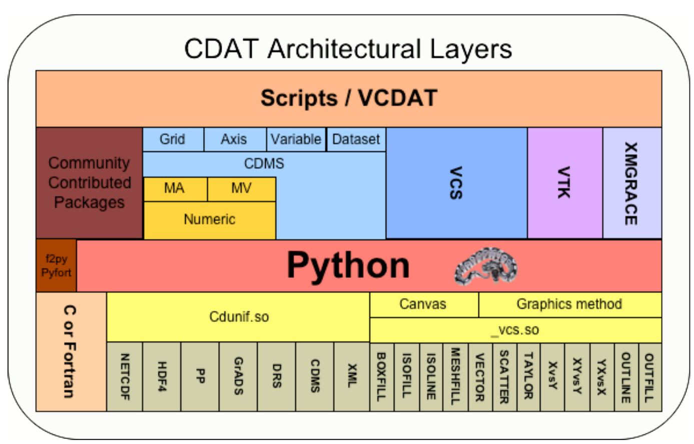
The Climate Data Analysis Tools (CDAT) team has prepared this report to describe the progressive development of the software. The CDAT team embodies an ongoing collaboration of climate and computer scientists in the Program for Climate Model Diagnosis and Intercomparison (PCMDI) (see **http://www-pcmdi.llnl.gov** and **http://cdat.sf.net**). This quarterly project report, which summarizes work on CDAT sponsored by the **U.S. DOE Office of Science** under the auspices of the **Office of Biological & Environmental Research** (OBER) program, elaborates CDAT's current objectives, scope, methods, and results, as well as mentioning additional relevant information, such as websites and collaborations. While aimed mainly at other specialists in climate modeling and analysis, these reports also provide information that is suitable for a wider scientific audience.

2. Background

For the past 1 ½ years, the CDAT development team has been working to deliver the latest version 4.0 of CDAT. We continue to address the challenges of enabling the management, discovery, access, and advanced analysis of the extensive datasets that are produced by climate and Earth system models. Well suited for combining enterprise applications and component assembly (i.e., graphical user interfaces and web services), CDAT has emerged as a next-generation analysis tool for climate scientists. Based on Python, a powerful user-friendly and widely used object-oriented scripting language, climate and associated computer scientists are provided with a significantly enhanced and productive working environment. CDAT development is organized around a multi-layer open-source design, which promotes collaboration between expert computer scientists at its core and research climate scientists as its end-users. The U.S. Department of Energy's **Program for Climate Model Diagnosis and Intercomparison** at Lawrence Livermore National Laboratory is central to the worldwide development and use of CDAT for gaining a better quantitative understanding of the Earth's climate system.

3. Design Philosophy

It is critical to build analysis tools that meet the goals of current and future climate and environmental studies. The CDAT architecture is the bridge between climate and environmental goals and a software-intensive system that must function in environments that typically involve interactions between software, hardware, people, data, physical spaces, organizational policies, standards, procedures, laws, and regulations. While quality attribute requirements (reliability, security, maintainability, etc.) drive the choice of software architecture, designing architecture that also satisfies functional requirements is vital to the success of the CDAT system.



We have based CDAT on the Python computing language because it has proved flexible and efficient for creating programs and scripts for climate applications. Many other groups also have adopted the Python open-source philosophy and are developing and sharing portable, flexible, and powerful software solutions in this framework, thereby rendering further CDAT development more cost effective. We anticipate that such inter-group collaborations will continue to make CDAT a better product for addressing key scientific questions.

The major CDAT components or subsystems are implemented as C modules. They are: the Climate Data Management System (CDMS – Cdunif.so) for manipulation of gridded data; Numerical Python (Numeric) for large-array mathematical operations; and the Visualization and Control System (VCS – vcs.so), Visualization Toolkit (VTK), and Xmgrace for computer graphics. In addition to these major modules, the software system includes various climate **community-contributed packages**, which are noted in the "CDAT Architectural Layers" figure shown above.

4. Data Management

The CDMS module (Climate Data Management System) provides the CDAT file I/O, data handling, regridding, and time representation capabilities. CDMS supports a uniform view of data that are ingested in different self-describing file formats, such as NetCDF, HDF, and GRIB. (ASCII and various sequential binary formats are handled by other CDAT modules.) The uniformity of I/O allows the writing of general analysis scripts that work correctly regardless of the underlying data format. At PCMDI, for example, data files are commonly kept in NetCDF or GRIB format, but users need not be aware of this detail. Moreover, CDMS provides the capability of combining hundreds of files into virtual *datasets*, which can be accessed as if they were a single file. This capability is essential for generalized processing of the high-volume, time-variant data produced by climate models.

In CDMS, the horizontal (latitude-longitude) coordinate structures of variables are represented on grids. The regrid module in CDMS allows the remapping of variables from one grid type or resolution to another, and can handle most grid types commonly encountered in climate models (e.g. rectilinear, curvilinear, and unstructured).

Finally, the cdtime module includes efficient functions for handling diverse calendars and time representations. For example, some models use a standard Gregorian or Julian calendar, while others omit leap years or adopt a 360-day year (i.e. universal 30-day months). Similarly, time values can be represented either as absolute, or as relative to a fixed base time.

PCMDI has been instrumental in developing and promoting acceptance of the Climate Format (CF) metadata standard. Based on the netCDF model of data, CF stipulates the writing of datasets that are fully self-describing, in the sense that each data variable has an associated description of its standard name attributes, physical units, and spatio-temporal coordinate structure. CDMS makes full use of such metadata, and produces CF-compliant data by default on output.

5. Numeric Processing

The computational functionality of CDAT is incorporated in three modules: Numeric, MA, and MV, where the first two of these were developed outside PCMDI. Numeric provides efficient array operations, including basic arithmetic and math functions. MA adds the ability to handle *masked arrays*, where the values of selected data elements are designated as missing or unknown—an especially common characteristic of observational datasets.

The MV module, which is native to CDMS, extends MA and Numeric by adding the concept of a *masked variable*, an array that has an associated coordinate structure. For example, if an MV array is sliced, the resultant data retains the corresponding time, latitude, and longitude boundaries. This capability makes it easy to interactively analyze datasets and to generate associated graphics with little effort.

6. Visualization

Visualization is an especially important capability for analysis of climate data. CDAT currently utilizes several visualization packages: the Visualization and Control System (VCS) – a 1D and 2D graphics system expressly designed to meet the needs of climate scientists; the Visualization ToolKit (VTK) – an open source 3D computer graphics, image processing, and visualization system; XmGrace – a 1D graphics package; and IaGraph – a VCS-based Python package for quick interactive graphing (see **http://www-pcmdi.llnl.gov/software-portal/cdat/tutorials/cdatbasics/ plotting-basics/**).

New to VCS in its current release is the ability to point and click on a plot to view quantitative information. Other VCS improvements include greater user-specified control over the picture template, graphics methods, and primitives; new graphics methods/projections and continental outlines; and more flexibility in plot aspect ratio and landscape/portrait orientation. The XmGrace module also was modified and enhanced to make it more consistent with VCS and easier to use.

For a future release, we are currently working with Ncvtk developers to make their 3D graphics accessible via CDAT. Ncvtk, which follows an event-driven model, is a collection of 3D visualization methods that can be interactively applied to data on structured lat/lon grids. However, for its initial beta release in mid-April 2006, the CDAT/Ncvtk module will not include geodesic or unstructured grids--a capability to be added later.

7. Front-end GUI and Scripting

The Visual Climate Data Analysis Tools (VCDAT, pronounced “v-c-dat”) is a graphical user interface (GUI) that allows CDAT capabilities to be used without requiring knowledge of the Python programming language. There have been numerous enhancements and additions to VCDAT since its prior official release. For help on using VCDAT, see **http://www-pcmdi.llnl.gov/software portal/cdat/ tutorials/ getting-started**.

Future VCDAT development will include the ability to directly access the data holdings of climate model simulations of the historical and climate-change scenarios specified by the Intergovernmental Panel on Climate Change (IPCC). Currently, these data holdings can only be accessed via the IPCC Earth Systems Grid (ESG) web portal (see details in Section 12 below). The planned enhancements of VCDAT will allow a user to select the IPCC ESG OPeNDAP server, authenticate access, list files and subdirectories, display metadata information, allow user-specified advanced searches, and download the chosen data directly into VCDAT for analysis. Server-side analysis also will be allowed in a later release of VCDAT.

Future VCDAT development also will include the further integration of Ncvtk. This effort will leverage the work already done by the Ncvtk developers in making their GUIs for 3D visualization methods accessible within VCDAT.

8. Web Services and Portal

A comprehensive collection of online CDAT documentation tutorials and examples are now available at the rapidly growing PCMDI website **http://www-pcmdi.llnl.gov/ software-portal/.** In addition to an expanding collection of standard manuals, there are now 15 tutorials covering both VCDAT GUI and CDAT scripting usage. The new "Tips and Tricks" section also provides useful advice for CDAT/Python programmers.

Future plans include the production of more advanced tutorials and an improved "News" section to provide more insight into our current areas of development. For developers and users interested in "getting under the hood", a CDAT API Reference also is now available.

The PCMDI Software Portal is built with the Python-based web content management system Plone, which enhances the potential for collaborative development of CDAT source code. Behind the scenes, we are migrating the CDAT source code repository to Subversion, an open-source revision control tool. We also plan to offer repository access and Software Portal accounts to outside developers, so that they can create and maintain CDAT Contrib Packages, as well as associated documentation. In addition, visitors will be able to view the repository (source code, revision history, etc.) directly at the website.

9. Scientific Use

CDAT was originally created primarily with the atmospheric and ocean modeling communities in mind, but it has now become a useful tool for the chemical/biogeochemical and other Earth system specialities as well. The range of application of CDAT varies from simple visualization of output from a single climate model, to navigation through many climate simulations, calculation and analysis of regional or global climate statistics, derivation and display of new quantities, and saving these in publication-quality form. CDAT enables users to read, analyze, display, and animate data in a variety of standard formats (e.g. NetCDF, HDF, GrADS/GRIB, PP, Binary, ASCII, and DRS), and makes it easy to write out data to a NetCDF file for future post-processing.

10. Broad Impact

ESG data holdings (see details in Section 12 below) enable scientists to analyze multiple climate model simulations from different centers, thereby furthering model intercomparison and diagnosis. For example, the current IPCC ESG data from numerous climate models is currently being assessed by hundreds of international researchers. In general, the benefits of such coordinated model intercomparison activities include increased communication among modeling groups, rapid identification and correction of gross modeling errors, the definition of standardized benchmarks, and a more complete and systematic record of modeling progress. The CDAT framework enables the international scientific community to easily view, access, and analyze these climate model data via either the VCDAT GUI interface or CDAT/Python scripts.

11. Collaboration

Where appropriate, CDAT development proceeds in collaboration with national and international groups (e.g., the British Atmospheric Data Centre, the NOAA Operational Model Archive and Distribution System (NOMADS) group, NSDL THREDDS, Global Organization for Earth System Science Portal (GO-ESSP), etc.). Collaboration areas include metadata schemas, web data portal design, OPeNDAP server, data storage, visualization, GUI development, climate analysis and diagnosis, and software infrastructure.

CDAT collaborations for this reporting period include:

- BADC – development of Data Extractor, VCDAT for Windows, PP format and data management
- University of Chicago – development of IaGraph and pyIoapi
- NOAA Geophysical Fluid Dyanamics Laboratory (GFDL) – development of Ncvtk

12. ESG Data Holdings

The Earth System Grid II (ESG) project, funded by the Department of Energy's Scientific Discovery through Advanced Computing (SCIDAC) program, has transformed climate model data into community resources. ESG has met this goal by creating a virtual collaborative environment that links climate centers and users around the world to models and data via a computing Grid, which is based on the Department of Energy's supercomputing resources and the Internet. ESG's success stems from partnerships between climate scientists and computer scientists to advance basic and applied research objectives.

In the Autumn of 2004, ESG began widespread distribution of the simulation data of more than 20 global climate models that were generated at the request of the Intergovernmental Panel on Climate Change (IPCC). The IPCC, which is jointly sponsored by the World Meteorological Organisation (WMO) and the

United Nations Environment Programme, carries out periodic assessments of the science of climate change. Fundamental to this effort is the production, collection and analysis of data from climate model simulations by major international centers. Analysis of a large set of standard numerical climate-change experiments (see above figure) and validating these against available observations provides more comprehensive understanding of the strengths and weaknesses of current climate models. The IPCC has requested PCMDI to collect model output data from these numerous simulations, and to efficiently distribute these to the community.

Model / Experiment	BCCR, Norway	CCCma, Canada	CCSR/NIES/FRCGC (hi-res), Japan	CCSR/NIES/FRCGC (med-res), Japan	CNRM, France	CSIRO, Australia	GFDL (CM2.0), USA	GFDL (CM2.1), USA	GISS (C4x3), USA	GISS (Model E-H), USA	GISS (Model E-R), USA	IAP, China	INM, Russia	IPSL, France	MPI, Germany	MRI, Japan	NCAR (CCSM3), USA	NCAR (PCM1), USA	NCC, China	UKMO (HadCM3), UK	UKMO (HadGEM1), UK
pre-industrial control	2	1	1	1	1	1	1	1	2	0	0	0	1	1	1	1	1	1	1	1	1
present-day control	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
climate of the 20th Century (20C3M)	2	1	1	3	1	1	1	1	2	0	0	0	1	1	1	5	5	4	8	2	2
committed climate change	2	1	0	1	1	1	1	0	0	0	0	0	0	1	1	1	5	4	0	1	1
SRES A2	0	3	0	3	1	1	1	1	0	0	0	0	1	2	1	5	5	4	4	1	1
720 ppm stabilization (SRES A1B)	2	3	1	3	1	1	1	1	2	0	0	0	1	1	1	6	5	4	4	1	1
550 ppm stabilization (SRES B1)	2	3	1	3	1	1	1	0	2	0	0	0	1	1	1	6	5	4	4	1	0
1%/year CO2 increase (to doubling)	2	1	1	3	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1
1%/year CO2 increase (to quadrupling)	0	1	0	3	1	0	1	0	0	0	0	0	1	1	1	1	1	1	1	0	0
slab ocean control	2	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1	0	0	1
2xCO2 equilibrium	2	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1	1	0	0	1
AMIP	0	1	1	3	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1	0	1

Simulations by Climate Modeling Group

The ESG project is now internationally recognized for this extensive distribution effort. Since December 2004, the IPCC model runs published into the IPCC ESG data holding have totaled over 26.6 TB (i.e., 60,300 files) and more than 500 scientific groups have registered to receive IPCC data for analysis. More than 335,000,files, equivalent to about 75 terabytes of data, have been downloaded, with the daily download rate averaging over 150 gigabytes. Some 250 research papers on analysis of the IPCC data also have been written thus far.

In current development work, the CDAT team is working with PyDAP, an open-source Python-based OPeNDAP server which will provide transparent data sub-setting and aggregation for interactive use. In the next release of CDAT, VCDAT users will be able to access remote files from the IPCC data portal as if they were stored locally, downloading only a user-selected subset of the data via OPeNDAP. PyDAP also will integrate with the existing IPCC Data Portal so that users will be able to download data either by means of FTP or OPeNDAP using the same username and password.

13. Summary

CDAT has evolved into a full-blown community-based system, with core developers applying cutting-edge computer science and technologies and with climate scientists exploring new diagnostic approaches in the CDAT framework. As an open-source project, CDAT has attracted many non-PCMDI contributors in continually developing its core and surrounding infrastructure (e.g., bug fixes, algorithms improvements, extensions, contrib packages, etc.). The international climate community also is increasingly adopting CDAT as a standard analysis tool, creating and sharing scripts for diagnosing diverse facets of climate model simulations.